TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Grado en Ingeniería Informática de Gestión y Sistemas de Información

ANÁLISIS Y VISUALIZACIÓN DE UN REPOSITORIO MULTICULTURAL: ESTUDIO DE LOS CONTENIDOS DE LAS DIFERENTES VERSIONES LINGÜÍSTICAS DE WIKIPEDIA

Memoria

ÁLVARO COSTA SÁNCHEZ
TUTORES: PERE BARBERAN AGUT
MARC MIQUEL I RIBÉ

2017 / 2018

TecnoCampus
Mataró-Maresme

Dedicatoria

Quiero dedicar este proyecto de final de carrera a mi familia, en concreto a mis padres y a mi hermano, que son las tres personas más importantes en mi vida.

Muchas gracias a todos, sin vosotros no hubiera llegado hasta aquí.

Resumen

Este proyecto pretende contribuir a la mejora continua de la enciclopedia online más grande del mundo aportando un estudio de su contenido, el cual se extrae de la plataforma que le da soporte. Éste contenido se analiza clasificándolo en un conjunto de categorías establecidas para poder analizar y visualizar su comportamiento y así determinar las preferencias de información de las distintas versiones lingüísticas existentes.

Resum

Aquest projecte pretén contribuir a la millora continua de l'enciclopèdia online més gran del mon aportant un estudi del seu contingut, el qual s'extreu de la plataforma que li dóna suport. Aquest contingut s'analitza classificant-lo en un conjunt de categories establertes per poder analitzar i visualitzar el seu comportament i així determinar les preferències d'informació de les diferents versions lingüístiques existents.

Abstract

This project aims to contribute to the continuous improvement of the largest online encyclopedia in the world providing a study of its content, which is extracted from the platform that supports it. The content is analysed by classifying it into a set of established categories in order to analyse and visualize its behaviour and thus determine the information preferences of the different existing language versions.

Índice

Índice de figuras

Índice de tablas

Glosario de términos

db	Data Base
GB	GigaByte
gz	GNU zip
JSON	JavaScript Object Notation
MB	MegaByte
SQL	Structured Query Language
TB	TeraByte
URL	Uniform Resource Locator
WWW	World Wide Web

1. Introducción

A lo largo de su existencia, la humanidad ha ido evolucionando de una manera muy rápida, pero en los últimos tiempos, este progreso ha superado cualquier límite establecido, y no hay indicios que lleven a pensar que se va a detener.

Esta evolución no sería posible sin la existencia de una base de conocimiento sólida que pueda dar respuesta a todas las preguntas y a todos los retos que se puedan presentar a lo largo del tiempo. Desde el origen de los tiempos, se ha ido transmitiendo de generación en generación, pero en muchos casos se ha podido perder información útil. Por ello, uno de los anhelos de la humanidad es poder unificar todo el conocimiento adquirido durante millones de años. Esto hecho serviría para que cualquiera pueda beneficiarse de esta información, y así seguir con el progreso.

En la época actual, denominada por muchos la sociedad del conocimiento, entra en juego un elemento completamente revolucionario, Internet. Es difícil entender la sociedad de hoy en día sin este elemento, ya que está completamente consolidado.

Internet se ha convertido en los últimos años en pieza clave en muchos aspectos, pero sobretodo se ha convertido en un componente capaz de unificar conocimiento e información.

Cualquier persona, independientemente de su geolocalización, puede acceder a cualquier tipo de información en cuestión de segundos. Basta con escribir la *URL* de una página web en un navegador o introducir una palabra o grupo de palabras en un buscador. En este último caso, se pueden obtener miles de resultados, pero en la mayoría de los casos, hay un resultado que siempre está presente, Wikipedia.

Cuando se habla de Wikipedia, se habla del repositorio online creado colaborativamente más grande de la humanidad y de la quinta página más consultada de Internet. Parte de la idea radical de dar acceso libre al conocimiento universal y tiene como objetivo dar una enciclopedia gratis y libre en todos los idiomas del planeta.

Este repositorio no se gestiona como un todo, sino que se divide en múltiples versiones a nivel mundial para poder cubrir el reto lingüístico, la cual cosa lleva a que cada país o región tenga una Wikipedia propia.

Este factor hace que el contenido que se publica sea rico y variado, y no esté sujeto solo a una única comunidad que proporciona la información, la cual puede ser de cualquier ámbito. Este hecho trae consigo grandes puntos positivos, como por ejemplo, que cada versión de Wikipedia sea independiente, o que la información que se facilita sea plural.

Pero lo que es una ventaja, a su vez puede convertirse en un inconveniente. Los términos que se alejan de temas técnicos, científicos o hechos puramente objetivos, están sujetos y condicionados por la cultura, región o país que ha aportado dicha información. Por este motivo, es difícil poder establecer una clasificación clara de todo el contenido que se proporciona y la vez también es complicado poder comprobar que temas son de interés para las diferentes comunidades lingüísticas.

Con este proyecto se pretende, por un lado, aprender y entender el funcionamiento, a nivel tecnológico, de Wikipedia y de Wikidata como base de datos que da soporte al repositorio colaborativo más grande del mundo, y por otro lado, analizar y visualizar el comportamiento del contenido en las distintas versiones existentes, clasificándolo en un conjunto de categorías establecidas para poder observar diferentes aspectos de significado y características relevantes de cada comunidad lingüística.

La principal motivación para la realización del proyecto es seguir mejorando esta gran comunidad sin ánimo de lucro, aportando un estudio de contenido de las diversas interpretaciones de Wikipedia, para de esta manera ver los intereses temáticos de cada comunidad y su preferencia de información, en base a datos objetivos existentes extraídos de la base de datos que da soporte a Wikipedia.

En los apartados posteriores se explican los puntos teóricos más importantes para entender el proyecto, los objetivos, los referentes, la metodología usada para llevar a cabo el proyecto, las partes más importantes del desarrollo realizado, los resultados obtenidos en el desarrollo y las conclusiones de los resultados.

2. Marco teórico

2.1. Contexto

La creación de Internet en la década de los ochenta permite en 1991 la puesta en marcha del *WWW* [1]. Se trata de un conjunto de protocolos que operan por encima de los protocolos de Internet y permiten un acceso flexible y generalizado a la información almacenada en la red de diversos formatos.

Este conjunto de protocolos da lugar al uso de un nuevo sistema de hipertexto para compartir documentos, que permite clasificar información de diversos tipos. El sistema es más conocido como *Web 1.0* [2], y es considerado como el acceso más sencillo y comprensible al universo de la información disponible en Internet. Se caracteriza por tener pocos productores de contenido, pero a la vez muchos lectores del mismo. Son páginas estáticas que no se actualizan de forma periódica.

Años más tarde se empieza a permitir el acceso a Internet a particulares y empresas, y desde ese mismo momento, la web se convierte en el servicio más empleado para ofrecer información de todo tipo. También aparecen diferentes tecnologías, tales como navegadores o lenguajes de programación enfocados en el entorno web y se perfecciona la manera de construir páginas web.

En 2004, se hace popular el término de *Web 2.0* [3]. Se trata de la evolución de la *Web 1.0*, en la que los usuarios dejan de ser meros espectadores y pasan a convertirse en contribuidores, siendo capaces de crear, dar soporte y formar parte de sociedades y/o comunidades tanto a nivel global como a nivel local. Se caracteriza por ser dinámica y por tener actualizaciones de forma regular. El término no es una tecnología en sí, sino una actitud en la que los usuarios toman para empezar a formar parte del contenido que circula por Internet.

La evolución de la web permite que aparezca un tipo de tecnología llamada *Wiki* [4]. Esta tecnología permite a múltiples usuarios combinar su información. Es un tipo de tecnología que incorpora una mentalidad hacia el logro de conocimiento.

Está diseñada para que los participantes puedan crear rápidamente nuevas páginas o editar páginas existentes. Representa una excelente plataforma tecnológica para la sociedad del conocimiento y se puede aprovechar para inspirar innovación y crecimiento.

2.2. Necesidades de información

2.2.1. Wikimedia Foundation

Wikimedia Foundation es la fundación, sin ánimo de lucro, que da soporte a Wikipedia. Su misión es fomentar el crecimiento, desarrollo y distribución de contenido educativo multilingüe y gratuito para que cualquier persona a nivel mundial tenga acceso al mismo [5].

Aparte de dar soporte a Wikipedia, también apoya otros proyectos que hacen uso de tecnologías wiki, tales como *Wiktionary* [6], diccionario de contenido en múltiples lenguajes; *Wikiquote* [7], repositorio de citas tomadas de personas famosas o personajes ilustres; *Wikibooks* [8], libros de texto, tutoriales, manuales u otros textos pedagógicos; *Wikisource* [9], biblioteca de textos originales que han sido publicados con una licencia de libre acceso; *Wikispecies* [10], repositorio de especies biológicas que tiene como objetivo abarcar todas las formas de vida que se conocen; *Wikinews* [11], fuente de noticias de todo tipo; *Wikiversity* [12], plataforma educativa donde se pueden crear proyectos de aprendizaje, crear contenido didáctico, entre otros; *Wikivoyage* [13], guía de viajes a nivel mundial; *Commons* [14], repositorio para archivos multimedia tales como imágenes, diagramas, videos, entre otros; *MediaWiki* [15], plataforma cuyo software es usado por todos los proyectos de Wikipedia y por otras Wikis que quieren seguir el mismo modelo; *Meta-wiki* [16], sitio dedicado a coordinar los proyectos de la fundación; *Incubator* [17], proyecto donde se puede desarrollar, escribir y probar nuevos idiomas y acciones para proyectos de la fundación; *Cloud Services* [18], ecosistema informático flexible que potencia la contribución técnica al mundo del software de Wikimedia; Wikidata.

Estos proyectos son el núcleo del movimiento Wikimedia. Son desarrollados en colaboración por usuarios de todo el mundo que utilizan software *MediaWiki*.

Todas las contribuciones se publican bajo una licencia *Creative Commons* [19] gratuita, lo cual hace posible que cualquier contenido sea utilizado libremente.

2.2.2. Wikipedia

Wikipedia es la enciclopedia de contenido libre más grande del mundo, la cual cualquiera puede leer y editar [20]. En 2017 fue la quinta página más visitada de Internet según un ranking de Alexa, compañía especializada en proporcionar datos y análisis comerciales de tráfico web [21].

Empieza como un proyecto en inglés el quince de Enero de 2001, a pesar de que actualmente tiene artículos redactados en doscientos noventa y tres idiomas [22]. Se funda bajo la creencia de que casi todo el mundo tiene algún conocimiento el cual pueda ser compartido con los demás.

Como proyecto colaborativo a nivel mundial, Wikipedia está sujeta a políticas y convenciones desarrolladas por la comunidad para describir las mejores prácticas, clarificar principios o resolver conflictos. No tiene reglas rígidas, pero se espera que la actitud de los editores sea acatar los principios establecidos en las políticas. Éstas se pueden editar, como cualquier página que se encuentre dentro del ámbito de Wikipedia [23].

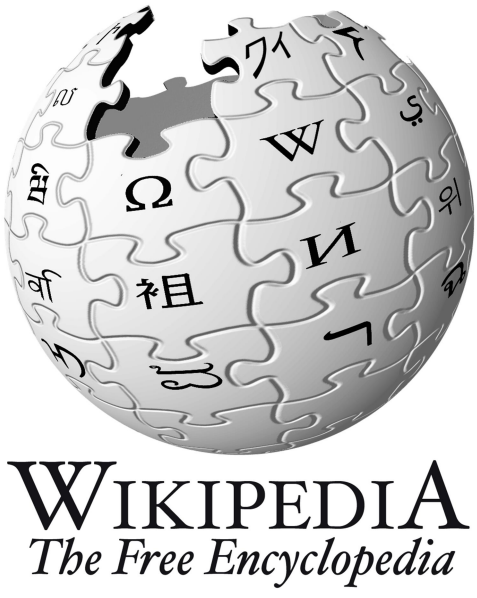

Fig. 2.2.2.1. Logo de Wikipedia. Fuente: Wikipedia, 2018.

2.2.3. Wikidata

Wikidata es la base de datos secundaria, colaborativa y multilingüe que da soporte a Wikipedia y a los demás proyectos relacionados que gestiona *Wikimedia Foundation* [25]. Tiene como objetivo proporcionar una fuente común para ciertos tipos de datos, como por ejemplo información de artículos de Wikipedia, fechas de nacimiento, entre otros.

El desarrollo inicial del proyecto está siendo supervisado por *Wikimedia Deutschland* y se divide en tres fases:

- Centralizar enlaces interlingüísticos.
- Proporcionar un lugar central para la información de las tablas de información de todas las versiones de Wikipedia.
- Crear y actualizar listados de artículos basados en datos de *Wikidata*.

Los datos que contiene están publicados bajo la licencia *Creative Commons Public Domain Dedication 1.0* [25], la cual permite la reutilización, la copia, la modificación, la distribución y la ejecución de su contenido, incluso con fines comerciales. Este contenido es proporcionado y mantenido por editores que trabajan sobre Wikidata.

Todos los datos de la base de datos son completamente multilingües, es decir, cualquier nuevo dato que se introduce está automáticamente en todos los lenguajes disponibles. Al ser datos estructurados, son fácilmente usables y cualquier usuario o software puede manipularlos.

Fig. 2.2.3.1. Logo de Wikidata. Fuente: Wikidata, 2018.

2.2.3.1. Estructura de los datos

Wikidata consta de ítems y de propiedades.

En el caso de los ítems, cada uno de ellos contiene un identificador, una descripción y cualquier cantidad de alias. El identificador se compone con una Q seguida de un número y son únicos.

En el caso de las propiedades, al igual que los ítems, también contienen un identificador y una descripción. Su identificador se compone con una P seguida de un número único.

Cada ítem posee un conjunto de propiedades que lo describen, y a su vez estas propiedades enlazan a un segundo ítem (valor), tal y como podemos ver en las Fig. 2.2.3.1.1. y 2.2.3.1.2.

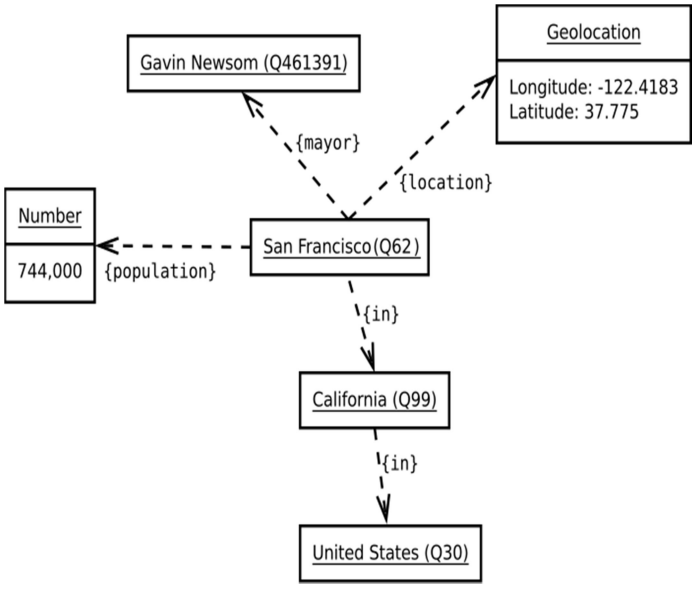

Fig. 2.2.3.1.1. Ejemplo de ítems y propiedades interconectados. Fuente: Wikidata, 2018.

Item	Property	Value
Q42	P69	Q691283
Douglas Adams	educated at	St John's College

Fig. 2.2.3.1.2. Ejemplo ítem – propiedad - valor. Fuente: Wikidata, 2018.

3. Objetivos y alcance

3.1. Objetivos

Este proyecto tiene dos grandes objetivos.

Por un lado, aprender y entender el funcionamiento de Wikipedia, y de Wikidata como base de datos que da soporte al repositorio colaborativo más grande del mundo. En este punto se hace especial hincapié en Wikidata, ya que es la plataforma donde se trabaja y donde se extraen todos los datos para llevar a cabo el estudio.

Por otro lado, analizar y visualizar el contenido de las distintas versiones de Wikipedia a nivel mundial, clasificándolo en un conjunto de categorías establecidas para poder observar diferentes aspectos de significado y características relevantes, para así determinar las preferencias de información de cada región o país.

En este objetivo es importante establecer un conjunto de categorías claras que permitan abarcar todos los temas, tomando como referencia estudios anteriores que tratan el tema del proyecto desde diferentes perspectivas. Estas categorías son un conjunto de macro grupos que agrupan el contenido de Wikidata en función del tema que tratan, como por ejemplo cultura, ciencia o deporte.

3.1.1. Acciones relacionadas con los objetivos principales

- Aprender cómo está clasificada la información en Wikipedia.
- Saber cómo funciona Wikidata y cómo es su estructura de datos interna.
- Definir las categorías para la clasificación del contenido.
- Clasificar el contenido en las categorías establecidas.
- Analizar el contenido una vez clasificado en categorías.
- Visualizar los intereses de información de las distintas versiones lingüísticas.
- Publicar el estudio de los contenidos para seguir aportando valor a todo el ecosistema de Wikipedia.

3.1.2. Producto y cliente

El producto resultante de la realización del estudio está muy enfocado al segundo gran objetivo del proyecto, es decir, el producto es el estudio y la investigación de los contenidos de Wikipedia a nivel global, proporcionando una visión objetiva, ya que se trabaja con datos reales almacenados en Wikidata.

Esta investigación se lleva a cabo por medio de distintas implementaciones que hacen posible toda la generación y extracción de los datos almacenados en Wikidata. Del mismo modo, una base de datos para almacenar todos los datos que se tratan y analizan. Estos dos componentes también forman parte del producto final, ya que sin ellos el estudio de los contenidos no es posible.

Por lo que respecta al perfil de cliente que va a utilizar el producto resultante del proyecto, destacar que el producto no se enfoca a un target concreto, ya que el estudio se realiza para que cualquier tipo de cliente pueda tomarlo como referencia para otros proyectos o simplemente para la obtención de un conocimiento objetivo y actual de los contenidos de las distintas versiones de Wikipedia existentes.

Es importante volver a recalcar que no existe un perfil concreto de cliente, ya que el estudio se realiza sobre plataformas que tienen un alcance global, muy conocidas y de fácil acceso desde cualquier dispositivo con conexión a Internet.

3.1.3. Objetivos del producto

- Extraer la información de Wikidata.
- Almacenar la información para que pueda ser analizada y visualizada.
- Visualizar de manera clara la información clasificada en las categorías definidas.
- Aportar un estudio de contenido a nivel mundial.
- Ayudar a la mejora continua de Wikipedia.

3.1.4. Objetivos del cliente

- Visualizar el estudio de los contenidos de las diferentes versiones de Wikipedia.
- Ver qué información interesa más en cada región.
- Cubrir cualquier necesidad de información relacionada con el estudio.
- Comparar el estudio con investigaciones anteriores.

3.2. Alcance

El proyecto queda establecido al cumplimiento de los objetivos marcados. Incluye la extracción, el procesamiento y la visualización de los datos obtenidos de los volcados de información oficiales de Wikidata. Estos volcados contienen toda la información existente en la plataforma.

Destacar que el estudio se realiza en base a unas categorías definidas, por lo tanto, el alcance del estudio viene dado por la clasificación de todo el contenido de Wikidata dentro de estas categorías.

Al tratarse de una plataforma conocida a nivel global y teniendo en cuenta que existen múltiples versiones lingüísticas, aunque el estudio recoja todo el contenido, se analizan resultados en base a muestras, es decir, a partir de un conjunto de versiones de Wikipedia se hacen las distintas comparativas del contenido clasificado. Este punto es importante ya que Wikipedia dispone de versiones en doscientos noventa y tres idiomas distintos.

4. Análisis de referentes

4.1. Estudio de Kittur

El estudio de Aniket Kittur se centra en el desarrollo de una técnica para mapear el contenido de Wikipedia utilizando su propia estructura de categorías. Se trata del primer mapeo cuantitativo de la distribución de temas en Wikipedia.

Defiende que existe contenido que es más fácil de clasificar que otro, ya que hay temas que interesan más, y explica que los fanáticos y la gente interesada en dichos temas ya se encarga de su clasificación. Al mismo tiempo, pide esfuerzos a la comunidad de Wikipedia para gestionar y clasificar el contenido de áreas específicas que no poseen de un interés suficiente para ser clasificado por terceros.

Opina que es necesario un método de clasificación automático, ya que el contenido de la plataforma ha crecido de manera exponencial en los últimos años. Hay que tener en cuenta que Wikipedia ya posee métodos de categorización, pero, según Kittur, son poco fiables.

El primer método se trata de un sistema de categorización de los artículos por medio de etiquetas que se encuentran en los mismos artículos. Éstos se pueden clasificar con múltiples etiquetas ya que puede tratar más de un tema a la vez. Según su visión, este sistema es difícil de entender y está mal formado, ya que cualquier usuario puede modificar las categorías de un artículo o introducir una nueva.

El segundo método de clasificación existente es un mecanismo de clasificación en forma de árbol. Éste sistema se utiliza en muchas wikis. Clasifica los artículos en las hojas del árbol, y los nodos superiores son las categorías. El problema de este mecanismo radica en que mientras la mayoría de categorías están distribuidas de manera equitativa, las categorías más generales tienen más del doble de nodos que las demás, cosa que hace inconsistente la clasificación del contenido. La clasificación por categorías de este método se realiza calculando la distancia entre los nodos, para de esta manera determinar los caminos más cortos.

Viendo, a su juicio, los fallos de los métodos de clasificación existentes, Kittur demuestra su técnica a partir de dos aplicaciones.

Por un lado, mapear la distribución de temas en Wikipedia y cómo han cambiado a lo largo del tiempo.

Por otro lado, mapear el grado de conflicto que se encuentra en cada una de las categorías que se establecen para realizar el estudio. En este punto destaca que la elección de las categorías aportar inconsistencia al estudio, ya que aunque las categorías están estructuradas de una manera jerárquica, no dejan de ser categorías establecidas subjetivamente, y puede llevar al hecho de que algunos elementos se queden sin clasificar.

A partir de la aplicación de la técnica mediante su algoritmo, el cual calcula los porcentajes a partir de la distancia que existe entre los diferentes nodos del árbol de categorías de Wikipedia, Kittur obtiene resultados que muestran que porcentaje de artículos existentes en Wikipedia pertenecen a cada una de las categorías que propone y cómo han evolucionado en el tiempo, tomando como referencia el contenido existente entre el año dos mil seis y el año dos mil ocho, tal y como podemos ver en el Fig. 4.1.1. Éste contenido fue extraído de Wikipedia y proporcionado por *Wikimedia Foundation* [26].

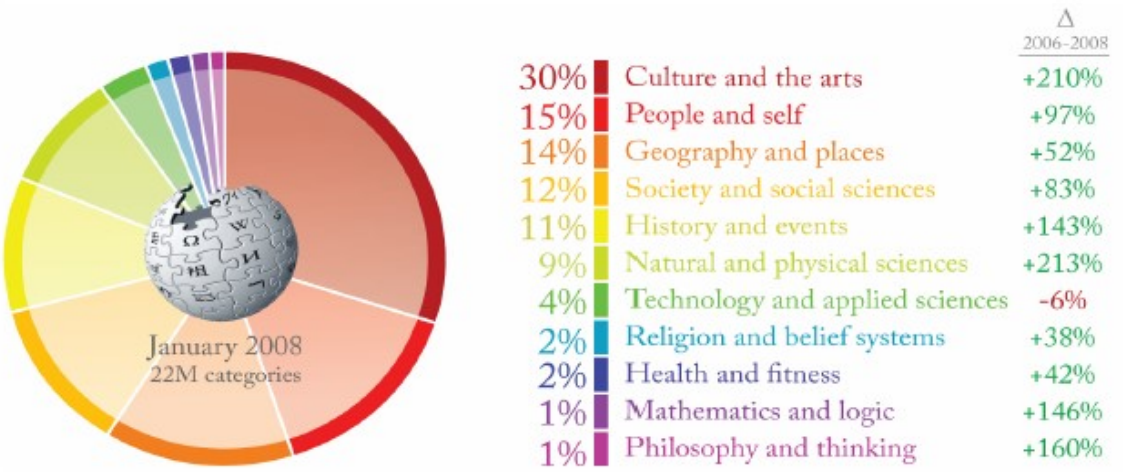

Fig. 4.1.1. Resultados del estudio de Kittur. Fuente: What's in Wikipedia?, 2018.

Una vez se obtienen los datos del contenido existente, Kittur analiza en porcentaje qué categorías poseen un mayor grado de conflicto en su distribución, es decir, para qué categorías es más difícil la asignación de artículos. Se puede observar ésta la información en la Fig. 4.1.2.

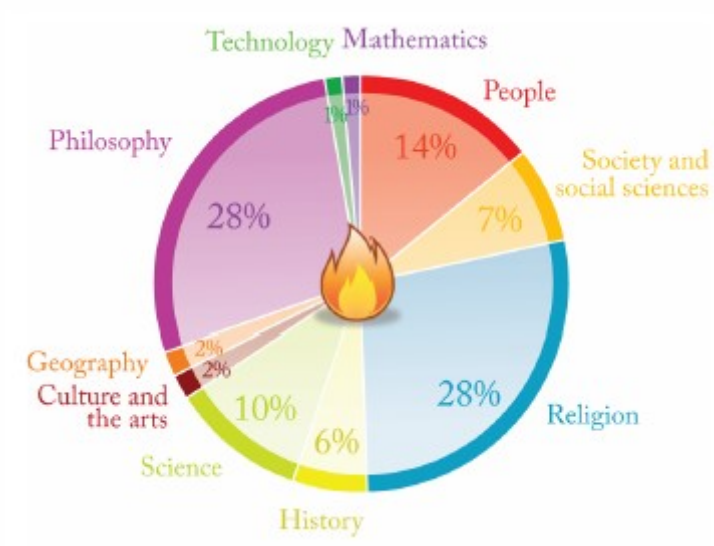

Fig. 4.1.2. Conflicto de las diferentes categorías de Kittur. Fuente: What's in Wikipedia?, 2018.

4.2. Estudio de Farina

El estudio de Jacopo Farina presenta una técnica que aprovecha el árbol de categorías existente en Wikipedia, el mismo que usa Kittur para su estudio, para asignar cada uno de sus artículos en una o más macro categorías.

Su estudio parte del algoritmo que usa Kittur, pero introduciendo y evaluando algunas variaciones. La más destacada es el cálculo de la similitud entre categorías según su coincidencia en artículos individuales.

Farina hace un enfoque del estudio bastante simple en su origen. Su idea es que si existen dos categorías conectadas por un mismo borde, es muy probable que dichas categorías estén relacionadas semánticamente.

Para la asignación de artículos a sus categorías, Farina hace una evaluación de las etiquetas que posee dicho artículo. Concretamente, calcula la proporción de las etiquetas de un artículo para de esta manera asignarlo a una de las macro categorías que el mismo establece.

En este punto se puede encontrar un matiz de incoherencia; un artículo puede pertenecer a dos macro categorías al mismo tiempo.

Farina lo solventa haciendo que la contribución de dicho artículo tenga el mismo peso en las distintas macro categorías a las cuales pueda pertenecer.

Además, introduce una variación más al estudio de Kittur. Introduce un factor de penalización. Este factor se aplica al hacer la asignación de una categoría existente en Wikipedia a una de las macro categorías que Farina propone.

Para la ejecución del estudio, Farina coge los datos de la versión inglesa de Wikipedia existentes en dos mil diez, y de nuevo, a diferencia de Kittur, Farina introduce dieciocho macro categorías en lugar de once. En la Fig. 4.2.1. se pueden ver el número de artículos en porcentaje, con el factor de penalización aplicado [27].

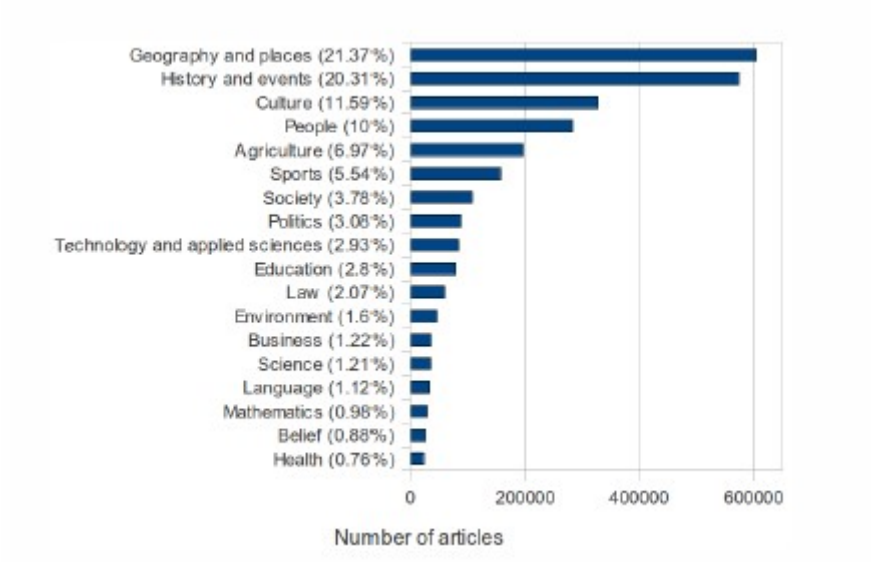

Fig. 4.2.1. Resultados del estudio de Farina. Fuente: Automatically assigning Wikipedia articles to macro-categories, 2018.

5. Metodología

5.1. Primer paso: Inmersión en el tema

Para la realización del proyecto se sigue una metodología combinada entre lo que es un estudio más científico y un proyecto informático. Pese a que el objetivo es realizar un estudio del contenido de Wikipedia, no se puede olvidar que para que el estudio sea objetivo se necesitan datos reales, lo cuales se consiguen por medio de procedimientos típicos de un proyecto de desarrollo.

Como en todo estudio o proyecto, lo primero es tener claro qué se quiere conseguir y sobre qué tema se está trabajando. En este caso, se pretende realizar un estudio de los contenidos de Wikipedia trabajando sobre la base de datos que le da soporte, Wikidata, de la cual se extrae toda la información necesaria para llevar a cabo el estudio. En este punto es importante tener una idea global de qué es Wikipedia, qué es Wikidata y cómo está estructurada su información.

5.2. Segundo paso: Definición de categorías

Al tratarse de un estudio de contenido y agregando que Wikipedia no tiene un sistema de categorización fiable, es necesario definir un conjunto de categorías para clasificar la información utilizada para el estudio. Se han seleccionado un total de diecinueve categorías, tomando como referencia los estudios de Kittur y Farina, para dar cabida a todos los artículos existentes en Wikipedia, información la cual se extrae de Wikidata. Son las siguientes:

Agriculture, Belief and Religion, Business, Culture and Arts, Education, Environment, Geography and Places, Food and Health, History and Events, Language, Law, Mathematics, People, Politics, Sciences, Society, Sports, Technology and Applied Sciences, Transport and Infrastructure.

5.3. Tercer paso: Obtención del contenido a analizar

Una vez situado el tema en cuestión y una vez seleccionadas las categorías, se necesitan los datos para poder llevar a cabo el estudio de contenido. Todos los datos se extraen de los volcados oficiales en formato JSON que tiene disponible Wikidata en la página oficial de la fundación Wikimedia.

Como el análisis del contenido de Wikipedia es en base a sus artículos, se necesita un volcado que contenga exclusivamente esa información. Además, también se necesita información sobre propiedades de dichos artículos, nombres, en cuantas lenguas están disponibles, entre otros. Por este motivo se necesita un segundo volcado de información. Hay que destacar que estos archivos vienen comprimidos con extensión .gz. El primer volcado tiene un peso en memoria de 103 MB y el segundo de 32 GB.

- *Primer volcado:* https://dumps.wikimedia.org/wikidatawiki/20180501/
  Archivo que contiene la lista de páginas con *main namespace*.
- *Segundo volcado:* https://dumps.wikimedia.org/wikidatawiki/entities/
  Directorio 20180507/wikidata-20180507-all.json.gz.

El primer volcado es el que contiene todos los identificadores de los ítems en Wikidata que referencian a artículos, y el segundo volcado contiene toda la información de la plataforma, en el cual se encuentran todos los ítems con su respectiva información, y todas las propiedades que describen a los mismos.

5.4. Cuarto paso: Estudio de la estructura del contenido

En este punto, cuando ya se tiene la información, hay que profundizar en cómo está estructurada la información dentro de los volcados y seleccionar sólo aquellos datos que sean útiles para el estudio. En el primero de ellos, la información viene en forma de lista, la cual contiene todos los artículos de Wikipedia. Éstos vienen con la notación que usa Wikidata, es decir, todos los ítems empiezan por una Q seguido de un identificador único. En el segundo volcado, cada ítem viene acompañado de más datos que se explican a continuación.

```
{
  "id": "Q60",
  "type": "item",
  "labels": {},
  "descriptions": {},
  "aliases": {},
  "claims": {},
  "sitelinks": {},
  "lastrevid": 195301613,
  "modified": "2015-02-10T12:42:02Z"
}
```


Fig. 5.4.1. Ejemplo de entidad dentro del volcado de información. Fuente: MediaWiki, 2018.

Como se puede apreciar en la Fig. 5.4.1., cada entidad dentro de Wikidata viene identificado por un identificador único, *id*. El campo *type* indica el tipo de la entidad. Es importante ya que en el mismo volcado de información se pueden encontrar entidades que no sean artículos, sino que sean propiedades que los definen. En este caso, el campo *id* empezaría por la letra P. En *claims* se encuentran todas las propiedades que posee dicho ítem.


```
{
  "claims": {
    "P17": [
      {
        "mainsnak": {
          "snaktype": "value",
          "property": "P17",
          "datatype": "wikibase-item",
          "datavalue": {
            "value": {
              "entity-type": "item",
              "numeric-id": 30
            },
            "type": "wikibase-entityid"
          }
        },
      },
```


Fig. 5.4.2. Ejemplo del campo claims de un ítem en Wikidata. Fuente: MediaWiki, 2018.

En la Fig. 5.4.2. se puede ver la estructura del campo *claims*. Se puede apreciar como cada propiedad posee un campo *mainsnak* que nos da información acerca ella. Concretamente se puede saber a qué hace referencia dicha propiedad. Las propiedades pueden referenciar otros ítems de Wikidata, como por ejemplo archivos tales como imágenes o videos, o incluso enlaces externos a páginas web de terceros. En este caso solo se necesitan aquellas propiedades que hagan referencia a otros ítems. Por este motivo es importante el campo *datatype*. Es el campo que determina a qué hace referencia la propiedad P.

Si la propiedad hace referencia a un ítem, es decir, a un artículo de Wikipedia, el campo *numeric-id* proporciona el identificador del ítem referenciado por la propiedad en cuestión. Solo se añade una Q en la parte izquierda del identificador para obtener el código del artículo en Wikidata.

Además del campo *claims*, también se necesita el campo *sitelinks*. Éste campo proporciona información acerca de los idiomas en los que está disponible el ítem. Además, como se puede apreciar en la Fig. 5.4.3., también proporciona el título del ítem en dicho idioma, en el campo *title*.


```
{
  "sitelinks": {
    "afwiki": {
      "site": "afwiki",
      "title": "New York Stad",
      "badges": []
    },
    "frwiki": {
      "site": "frwiki",
      "title": "New York City",
      "badges": []
    },
```


Fig. 5.4.3. Ejemplo del campo sitelinks de un ítem en Wikidata. Fuente: MediaWiki, 2018.

5.5. Quinto paso: Creación de la estructura de almacenamiento

Llegado a este punto, con los volcados de contenido procedentes de Wikidata y con el conocimiento de qué datos se necesitan para el estudio del contenido, se procede a la creación de la base de datos, con diferentes tablas, para almacenar toda la información y que ésta sea fácilmente accesible en cualquier momento.

Es necesaria la creación de siete tablas en la base de datos para albergar toda la información extraída de los volcados de contenido de Wikidata. En la Tabla 5.5.1. se puede ver qué se almacena en cada una de ellas.

Tablas	Contenido
Primera	Identificadores de los ítems que referencian a otros ítems.
Segunda	Identificadores de las propiedades y sus nombres.
Tercera	Tripletas ítem – propiedad – ítem.
Cuarta	Identificadores de los ítems, idioma en los cuáles está disponible y título en dicho idioma.
Quinta	Coeficientes de las propiedades no neutras clasificadas en las distintas categorías.
Sexta	Coeficientes de las propiedades neutras más los coeficientes clasificados en las categorías.
Séptima	Coeficientes finales normalizados a uno de las distintas categorías.

Tabla 5.5.1. Breve explicación de la información almacenada en la base de datos. Fuente: Elaboración propia, 2018.

En la primera tabla se almacenan los identificadores extraídos del primer volcado de información. Ésta información es necesaria para poder filtrar la información que hace referencia a artículos en el segundo volcado.

La segunda tabla contiene los identificadores y los nombres de las propiedades. Se extraen del segundo volcado de información.

La tercera contiene toda la información que se usa más tarde para poder calcular los coeficientes necesarios que determinan la cantidad de información asociada a cada una de las categorías que se definen para este proyecto. Esta información se estructura en forma de tripletas. Para cada ítem analizado se guarda su identificador junto a todas las propiedades que hacen referencia a artículos. Además, también se guardan los valores referenciados por cada propiedad. Toda la información de esta tabla se extrae del segundo volcado de información.

En la cuarta tabla se almacena la información relacionada con los idiomas en los que está disponible cada ítem analizado y sus respectivos títulos. Se extrae del segundo volcado de información.

En la quinta y sexta tabla se almacena los coeficientes calculados en las dos primeras iteraciones del proceso para este cometido. Los coeficientes se calculan a partir de la tabla que contiene la información de las tripletas.

En la séptima tabla se guardan los coeficientes finales que se usan para la visualización de los resultados obtenidos en el proyecto.

El proceso de cálculo de coeficientes se explica con más detalle en apartados posteriores.

5.6. Sexto paso: Extracción de información

En este paso se procede a la manipulación del contenido de los volcados para poder hacer posible la extracción de la información relevante para realizar el estudio.

En el primer paso se realiza la extracción de la información del primer volcado y se almacena en la primera de las tablas de la base de datos destinada a contener exclusivamente los identificadores de todos los ítems de Wikidata que hacen referencia a artículos.

En el segundo paso se accede al segundo volcado y se almacenan todas las propiedades existentes en Wikidata con su respectivo nombre. Esta información se guarda en la segunda tabla de la base de datos.

En el tercer paso se realiza una extracción de datos más exhaustiva. Se debe acceder y se debe recorrer el segundo volcado de información y también se debe almacenar la información de las tripletas ítem – propiedad – ítem referenciado en su respectiva tabla de la base de datos. Al mismo tiempo, también se debe guardar, en otra tabla, la información que hace referencia a los ítems, los lenguajes en los que ese ítem está presente y su respectivo nombre en cada uno de ellos.

5.7. Séptimo paso: Análisis y clasificación de propiedades

Una vez que la información necesaria está almacenada en la base de datos, se puede pasar a la primera fase de análisis. En esta fase se debe establecer qué propiedades se van a tomar como referencia para realizar el estudio del contenido. Para hacerlo, se tiene que comprobar que propiedades tienen más ocurrencias en Wikidata, para de esta manera cubrir el mayor porcentaje de información posible.

Nº de propiedades	Nº de ocurrencias	Porcentaje (%)
1032	**175.590.911**	**100**
200	**174.381.782**	**99,31**

Tabla 5.7.1. Tabla de ocurrencias de propiedades. Fuente: Elaboración propia, 2018.

Con la información almacenada, teniendo en cuenta que solo se tiene la información de los ítems, es decir, artículos, si se consulta en la base de datos se obtiene la información de las propiedades con sus ocurrencias en la tabla de tripletas. Como se puede ver en la Tabla 5.7.1., cogiendo las doscientas primeras propiedades ordenadas por número de ocurrencias, se obtiene un porcentaje muy elevado.

En Wikidata existen cuatro mil quinientas setenta y ocho propiedades, pero solo mil treinta y dos hacen referencia a ítems que tienen relación con artículos en Wikipedia. El número de ocurrencias se calcula contando todas las apariciones de una misma propiedad en la tabla de tripletas. Por ejemplo, la propiedad *located in the administrative territorial entity* tiene más de cinco millones y medio de ocurrencias dentro de la tabla de tripletas, mientras que la propiedad *symbolizes* tiene solo dos ocurrencias. En este caso y como indica la tabla previamente mencionada se cogen las doscientas primeras propiedades con más ocurrencias, y se descartan todas las demás.

Las doscientas propiedades con más ocurrencias se almacenan manualmente en diecinueve estructuras de datos, una para cada una de las distintas categorías utilizadas para el estudio del contenido.

Pero no es todo, en este punto se puede encontrar un pequeño problema. En Wikidata existen propiedades P que no acaban de describir al elemento que ésta referencia. Por ejemplo, *Nelson Mandela* (Q8023) – *instance of* (P31) – *human* (Q5). En este caso, la propiedad *instance of* no tiene un valor semántico que sirva para describir al ítem Nelson Mandela. Lo que verdaderamente lo describe, es el ítem referenciado, human, no la propiedad en sí. Estos casos hay que tratarlos de manera especial ya que estas propiedades no se pueden clasificar en las categorías definidas. Son propiedades neutras. Por este motivo se necesita otra estructura de datos para almacenar estas propiedades.

Nº de Propiedades descriptivas	Nº de Propiedades neutras
171	**29**

Tabla 5.7.2. Tabla de tipos de propiedades. Fuente: Elaboración propia, 2018.

Como se puede ver en la Tabla. 5.7.2., entre las doscientas propiedades con más ocurrencias, veintinueve no describen al ítem, es decir, son propiedades neutras. Las ciento setenta y una propiedades descriptivas se clasifican en las distintas estructuras de datos definidas para cada una de las categorías. Las veintinueve propiedades restantes se clasifican en la estructura de datos destinada a las propiedades neutras o no descriptivas.

5.8. Octavo paso: Cálculo de coeficientes

Una vez que la información necesaria está en la base de datos y una vez que las propiedades, tanto las que se clasifican en las categorías definidas como las neutras, están clasificadas en las estructuras de datos, se puede proceder al cálculo de los coeficientes para determinar qué categorías poseen más información.

Para este paso son necesarias tres iteraciones para obtener los coeficientes finales.

En la primera iteración, se recorre el segundo volcado que contiene toda la información de Wikidata. Para cada ítem Q del volcado se comprueba que éste aparezca en la tabla dónde se almacenan todos los identificadores de los ítems, es decir, la primera de las tablas definidas en la Tabla 5.5.1. Si el ítem aparece en esta tabla, se miran todas sus propiedades, cogiendo solo aquellas que hagan referencia a otros ítems, es decir, que su campo *datatype* sea igual a *wikibase-item*.

Una vez se consiguen todas las propiedades del ítem en cuestión, se comprueba qué propiedades aparecen en las distintas estructuras de datos pertenecientes a las categorías definidas anteriormente.

En este punto se destaca que no se tienen en cuenta las propiedades clasificadas como neutras para el cálculo de los coeficientes en este paso, pero sí que se tienen en cuenta para hacer la ponderación. Por ejemplo, si el ítem Q posee treinta propiedades, veintiocho descriptivas y dos neutras, se tienen en cuenta las treinta propiedades pero solo se calculan los coeficientes de las veintiocho propiedades clasificadas en categorías. Es decir, si de esas treinta, dos pertenecen a la categoría *sports*, se calcula el coeficiente del ítem para esa categoría dividiendo dos entre treinta. De esta manera la suma de coeficientes de todas las categorías pertenecientes al ítem Q no será uno.

En la segunda iteración, también se recorre todo el volcado que contiene toda la información de Wikidata y se comprueba que el identificador del ítem Q en cuestión esté en la tabla de identificadores. La diferencia con la primera iteración es que ahora sólo se miran las propiedades definidas como neutras. Sola se cogen las propiedades que existan en la estructura de datos definida para las propiedades neutras o no descriptivas.

Aparte de las propiedades, también se cogen los ítems Q referenciados por cada propiedad neutra. De cada ítem Q referenciado, se recuperan los coeficientes calculados en la primera iteración. Éstos se deben almacenar para no perder dicha información. Después, los coeficientes de la primera iteración se deben ponderar sobre el mismo valor utilizado en la primera iteración, treinta en el ejemplo previamente explicado, que es el peso que tiene la propiedad neutra dentro del ítem Q.

Una vez calculados todos los nuevos coeficientes, se suman todos los coeficientes de cada una de las categorías, es decir, se suman los coeficientes de cada ítem Q obtenidos en la primera iteración con los coeficientes obtenidos en ésta segunda iteración.

En la tercera iteración, se hace exactamente lo mismo que en la segunda iteración, con la diferencia que esta vez para cada ítem Q referenciado por las propiedades neutras, se cogen los coeficientes calculados en la segunda iteración, y no en la primera.

De esta manera, los coeficientes de cada ítem Q de Wikidata de las distintas categorías definidas sumarán uno o cero coma noventa y nueve.

5.9. Noveno paso: Visualización de la información clasificada

Una vez calculados y almacenados todos los coeficientes de las categorías definidas para el estudio de todos los ítems Q de Wikidata, se procede a la visualización de los resultados.

Para ello es necesaria una implementación que sea capaz de coger los coeficientes obtenidos en el paso anterior, y a partir de ellos, generar gráficos que muestren los porcentajes de las distintas categorías seleccionadas para el estudio.

Destacar que para cada categoría se suma el coeficiente obtenido en todos los elementos analizados, y después se pondera por la suma total de todos los coeficientes para obtener un valor ponderado a uno.

Mediante el uso de una librería específica para la generación de gráficos se generan los distintos elementos para que puedan ser analizados.

5.10. Décimo paso: Análisis de la visualización

En este paso se procede a hacer un análisis profundo de los resultados obtenidos durante todo el proceso.

Destacar que el análisis se hace en base a los resultados obtenidos sobre los datos globales, es decir, el contenido de todas las versiones existentes agrupado en las distintas categorías definidas para el estudio, sobre la versión más antigua y con más contenido de Wikipedia, la versión inglesa, y sobre las versiones en español y en catalán.

6. Desarrollo

6.1. Herramientas y tecnologías

Para la realización del proyecto se han usado distintas herramientas y tecnologías *open-source*.

Para la base de datos se usa el sistema de gestión SQLite [28]. Cabe remarcar que SQLite es la mejor opción en comparación a otros sistemas de gestión, como por ejemplo MySQL, ya que los datos extraidos de los volcados se almacenan en un archivo con extensión *.db*, portable, fácilmente gestionable y que no requiere de ninguna instalación pesada. Además, se pueden crear bases de datos de hasta 2 TB, lo cual para lo que a este estudio respecta es suficiente.

- SQLite binario pre compilado para Windows – versión 3.23.1.
- SQLite Studio para Windows – versión 3.1.1.

En cuanto a lenguaje de programación, se ha optado por Python [29]. Para la parte de desarrollo del proyecto es la mejor opción ya que la implementación se basa en scripts independientes y facilita las tareas de extracción de datos. Otro motivo para utilizar Python en lugar de otros lenguajes de programación como por ejemplo Java o C++ es que muchos de los proyectos existentes que trabajan sobre Wikidata están implementados en Python.

- *Release* de Python para Windows - versión 2.7.14.

Por lo que al entorno de programación respecta, se ha decidido usar Eclipse [30] para desarrolladores Java, instalando un *plug-in* para Python, PyDev [31].

- Eclipse IDE para desarrolladores Java – versión neon 3.
- *Plug-in* PyDev para Eclipse.

Para la parte de visualización de datos, se usa la librearía de Python Bokeh [32].

- Librería de Python Bokeh – versión 0.12.16.

6.2. Decisiones de implementación

6.2.1. Creación de la estructura de almacenamiento

Teniendo claros los cuatro primeros pasos de la metodología empleada para la realización del proyecto, se empieza el desarrollo para obtener los resultados del estudio de los contenidos de Wikipedia extraídos de Wikidata.

En primer lugar, y haciendo referencia al quinto paso de la metodología, se crea la base de datos con sus respectivas tablas. El uso de SQLite facilita mucho el trabajo, basta con crear un archivo y otorgarle una extensión .db.

Para la creación de las tablas, basta con abrir SQLite Studio, apartado *Tools*, y seleccionar la opción *Open SQL editor*.

En primer lugar, se crea la tabla que guarda todos los ítems de Wikidata que son artículos en Wikipedia. Esta tabla solamente contiene un campo llamado *q_item*, que a su vez es clave primaria, tal y como se puede observar en la Fig. 6.2.1.1.

```
CREATE TABLE `Item` (
    `q_item`    TEXT NOT NULL,
    PRIMARY KEY(`q_item`)
);
```

Fig. 6.2.1.1. Código para la creación de la tabla Item. Fuente: Elaboración propia, 2018.

En segundo lugar, se crea la tabla que almacena todas las propiedades existentes en Wikidata junto con sus nombres. Esta tabla se compone de dos campos: el primero *p_property*, que hace referencia al identificador de la propiedad, y segundo *p_name*, que hace referencia al nombre de dicha propiedad. Aprovechando que *p_property* es un identificador único, se usa como clave primaria, tal y como se puede ver en la Fig. 6.2.1.2.

```
CREATE TABLE `Property` (
    `p_property`    TEXT NOT NULL,
    `p_name`    TEXT NOT NULL,
    PRIMARY KEY(`p_property`)
);
```

Fig. 6.2.1.2. Código para la creación de la tabla Property. Fuente: Elaboración propia, 2018.

En tercer lugar, se crea la tabla que debe almacenar las tripletas Q1 – P – Q2, es decir, ítem – propiedad – ítem referenciado. Esta tabla se compone de cuatro campos: *id*, hace referencia al identificador de la tripleta; *q_item*, hace referencia al identificador del ítem; *p_property*, indica el identificador de la propiedad del ítem y *q2_item*, que hace referencia al identificador del ítem referenciado por la propiedad. Se puede ver el código SQL de la creación de la tabla en la Fig. 6.2.1.3.

```
CREATE TABLE `Triplets` (
    `id`    INTEGER NOT NULL,
    `q_item`    TEXT NOT NULL,
    `p_property`    TEXT NOT NULL,
    `q2_item`   TEXT NOT NULL,
    PRIMARY KEY(`id`)
);
```

Fig. 6.2.1.3. Código para la creación de la tabla Triplets. Fuente: Elaboración propia, 2018.

En cuarto lugar, es necesaria la creación de una tabla para almacenar los idiomas en los cuales se puede encontrar un ítem determinado, y el título de los mismos en la lengua en cuestión. La tabla se compone de tres campos: *q_item*, identificador del ítem de Wikidata; *site_lang*; código del idioma del ítem, y *q_title*, título del ítem en ese idioma. La clave primaria de la tabla la componen el ítem y el código del idioma como se puede apreciar en la Fig. 6.2.1.4.

```
CREATE TABLE `Sitelink` (
    `q_item`    TEXT NOT NULL,
    `site_lang` TEXT NOT NULL,
    `q_title`   TEXT NOT NULL,
    PRIMARY KEY(`q_item`,`site_lang`)
);
```

Fig. 6.2.1.4. Código para la creación de la tabla Sitelink. Fuente: Elaboración propia, 2018.

En quinto lugar, se precisa de la creación de una tabla en la cual se almacenan los coeficientes calculados en la primera iteración. Esta tabla se compone por el ítem poseedor de los coeficientes, *q_item*, el número de apariciones de las propiedades en las estructuras donde se almacenan las propiedades clasificadas, *n_properties*, y todos los coeficientes para cada una de las categorías definidas. En la Fig. 6.2.1.5. se puede ver el código SQL utilizado para la creación de la tabla.

```
CREATE TABLE `Coef_First_Step` (
    `q_item`    TEXT NOT NULL,
    `n_properties`  INTEGER NOT NULL,
    `agriculture`   REAL NOT NULL,
    `belief`    REAL NOT NULL,
    `business`  REAL NOT NULL,
    `culture`   REAL NOT NULL,
    `education` REAL NOT NULL,
    `environment`   REAL NOT NULL,
    `geography` REAL NOT NULL,
    `health`    REAL NOT NULL,
    `history`   REAL NOT NULL,
    `language`  REAL NOT NULL,
    `law`   REAL NOT NULL,
    `mathematics`   REAL NOT NULL,
    `people`    REAL NOT NULL,
    `politics`  REAL NOT NULL,
    `sciences`  REAL NOT NULL,
    `society`   REAL NOT NULL,
    `sports`    REAL NOT NULL,
    `technology`    REAL NOT NULL,
    `transport` REAL NOT NULL,
    PRIMARY KEY(`q_item`)
);
```

Fig. 6.2.1.5. Código para la creación de la tabla Coef_First_Step. Fuente: Elaboración propia, 2018.

En sexto lugar, se necesita crear la tabla en la cual se almacenan los coeficientes generados por la segunda iteración del proceso de cálculo de coeficientes. Esta tabla es exactamente igual a la tabla anterior. Debe ser así ya que se recalculan los coeficientes cogiendo la información de la tabla *Coef_First_Step* y los nuevos coeficientes calculados en la segunda iteración en base a las propiedades neutras encontradas en cada uno de los ítems del volcado.

En séptimo y último lugar, es necesaria la creación de la tabla para almacenar los coeficientes finales, los cuales se van a usar para la visualización de los resultados. En esta tabla se almacenan los coeficientes resultantes de la tercera iteración del proceso, y que se usan para la visualización de los resultados.

Sigue la misma línea que las dos tablas anteriores, pero sin incluir el campo *n_properties*. En la Fig. 6.2.1.6. se puede ver el código para la creación de la misma.

```
CREATE TABLE `Coefficient` (
    `q_item`    TEXT NOT NULL,
    `agriculture`   REAL NOT NULL,
    `belief`    REAL NOT NULL,
    `business`  REAL NOT NULL,
    `culture`   REAL NOT NULL,
    `education` REAL NOT NULL,
    `environment`   REAL NOT NULL,
    `geography` REAL NOT NULL,
    `health`    REAL NOT NULL,
    `history`   REAL NOT NULL,
    `language`  REAL NOT NULL,
    `law`   REAL NOT NULL,
    `mathematics`   REAL NOT NULL,
    `people`    REAL NOT NULL,
    `politics`  REAL NOT NULL,
    `sciences`  REAL NOT NULL,
    `society`   REAL NOT NULL,
    `sports`    REAL NOT NULL,
    `technology`    REAL NOT NULL,
    `transport` REAL NOT NULL
);
```

Fig. 6.2.1.6. Código para la creación de la tabla Coefficient. Fuente: Elaboración propia, 2018.

6.2.2. Extracción de información

En este punto es necesaria la implementación de tres scripts que sean capaces de acceder a los volcados comprimidos de información de Wikidata, y que al mismo tiempo almacenen la información relevante para el estudio en la base de datos previamente creada.

El primer script se encarga de almacenar los identificadores de todos los ítems de Wikidata que referencian a artículos. Para ello, es necesaria la conexión a la base de datos creada previamente. En la Fig. 6.2.2.1. se puede ver el código Python para dicha conexión.

```
# SQLite path file
sqlite_file = 'D:\WikidataDB\wikidata.db'
# Connect to data base
connection = SQLite.connect(sqlite_file)
# Cursor in order to execute the queries
cursor = connection.cursor()
```

Fig. 6.2.2.1. Código para la conexión a la base de datos. Fuente: Elaboración propia, 2018.

También es necesario acceder y recorrer el volcado que contiene la información de los identificadores que hacen referencia a artículos. En la Fig. 6.2.2.2. se puede una parte del código empleado para esta tarea.

```
# JSON dump path file
dump_path = 'D:\WikidataDB\dumps\wikidatawiki-20180501-all-titles-in-ns0.gz'
dump_in = GzipFile(dump_path, 'r')
# Read the first line ('page_title')
line = dump_in.readline();
# Iterate the dump
line_num = 0
while line != '':
    # Read dump line
    line = dump_in.readline()
```

Fig. 6.2.2.2. Fragmento de código para el acceso del volcado de información. Fuente: Elaboración propia, 2018.

Con el accedo a la base de datos y el volcado, se procede a almacenar la información. Esta información se almacena en la tabla *Item* previamente creada. En la Fig. 6.2.2.3. se puede observar parte de este proceso.

```
# Split the line
line_array = line.split('\n')
try:
    # Q item
    qitem = line_array[0]
    # Insert query
    sql_query = 'INSERT INTO Item (q_item) VALUES(?)'
    cursor.execute(sql_query, (qitem, ))
    # Commit when a thousand rows are processed
    if line_num == 1000:
        # Reset the line_number to 0
        line_num = 0
        # Save the changes to the database
        connection.commit()
```

Fig. 6.2.2.3. Fragmento de código para la inserción en la base de datos. Fuente: Elaboración propia, 2018.

El segundo script se centra en almacenar la información de todas las propiedades existentes en Wikidata. Para ello es necesaria la conexión a la base de datos y el acceso al volcado. Esta vez el volcado es el que contiene toda la información, no solo los identificadores de los ítems. En la Fig. 6.2.2.4. se puede ver parte del proceso de la obtención de propiedades.

```
# Parse line into JSON
ent = json.loads(line.rstrip('\n,'))
# Find properties
if ent['type'] == 'property':
    # Property name
    p_property = ent['id']
    # Take labels
    labels = ent['labels']
    # Property name
    p_name = labels['en']['value']
    # Insert query
    sql_query = 'INSERT INTO Property (p_property, p_name) VALUES(?, ?)'
    cursor.execute(sql_query, (p_property, p_name))
```

Fig. 6.2.2.4. Fragmento de código del proceso de obtención de propiedades. Fuente: Elaboración propia, 2018.

En el tercero, se aborda el tema de las tripletas de información y del contenido lingüístico de cada ítem Q.

En la primera fase, se analiza y se extrae toda la información referente a las tripletas. En la Fig. 6.2.2.5. se puede ver parte del proceso para esta primera fase de análisis. Se cogen las propiedades del ítem Q analizado que hacen referencia a otros ítems y se almacenan en base de datos las tripletas ítem – propiedad – ítem referenciado.

```
# Property
p_property = mainsnak['property']
# Check if property references Q item
if mainsnak['datavalue']['type'] == 'wikibase-entityid':
    # Increase the ID value
    id = id + 1
    # Q2 custom_value
    q2item = 'Q{}'.format(mainsnak['datavalue']['value']['numeric-id'])
    # Insert query
    sql_query = 'INSERT INTO Triplets (id, q_item, p_property, q2_item)
    cursor.execute(sql_query, (id, q_item, p_property, q2item))
```

Fig. 6.2.2.5. Fragmento de código para el tratamiento de las tripletas. Fuente: Elaboración propia, 2018.

En la segunda fase, se trata el tema lingüístico, y se obtiene para cada ítem, en qué lenguas aparece y el título en dicha lengua. En la Fig. 6.2.2.6. se puede apreciar parte del código empleado para este proceso.

```
# Sitelinks
sitelinks = ent['sitelinks'].items()
# Iterate over sitelinks
for key, value in sitelinks:
    item_array = key.split('wiki')
    # Avoid Wikipedia German projects
    if len(item_array) == 2 and item_array[1] == '':
        # Language code
        site = item_array[0]
        # Title
        title = value['title']
        # Insert query
        sql_query = 'INSERT INTO Sitelink (q_item, site_lang, q_title)
        cursor.execute(sql_query, (q_item, site, title))
```

Fig. 6.2.2.6. Fragmento de código para el tratamiento de los códigos lingüísticos. Fuente: Elaboración propia, 2018.

6.2.3. Análisis y clasificación de propiedades

Para el análisis y clasificación de propiedades, se crean veinte estructuras de datos para dar cabida a las diecinueve categorías usadas para el estudio y para las propiedades neutras o no descriptivas.

En este paso del desarrollo, se obtienen las doscientas categorías con más ocurrencias en Wikidata. Para ello, se realiza una consulta SQL a la base de datos creada, en concreto a la tabla *Triplets* unida a la tabla *Property*. En la Fig. 6.2.3.1. se puede ver el código SQL de la consulta realizada.

```
SELECT t.p_property AS Property,
       p.p_name AS Name,
       COUNT(t.p_property) AS Ocurrences
FROM Triplets t INNER JOIN Property p
     ON (t.p_property = p.p_property)
GROUP BY t.p_property
ORDER BY 3 DESC
```

Fig. 6.2.3.1. Consulta para obtener las propiedades con más ocurrencias. Fuente: Elaboración propia, 2018.

En la Fig. 6.2.3.2. se pueden ver las diez primeras propiedades resultantes de ejecutar el código de la consulta que se muestra en la Fig. 6.2.3.1.

#	Property	Name	Ocurrences
1	P31	instance of	43888968
2	P2860	cites	36672747
3	P1433	published in	16372917
4	P17	country	8727981
5	P131	located in the administrative territorial entity	5592645
6	P106	occupation	3771811
7	P21	sex or gender	3717661
8	P27	country of citizenship	2837629
9	P735	given name	2630188
10	P171	parent taxon	2484203

Fig. 6.2.3.2. Tabla con las diez primeras propiedades con más ocurrencias. Fuente: Elaboración propia, 2018.

Una vez obtenidas las propiedades, se clasifican manualmente en las veinte estructuras de datos creadas. En la Fig. 6.2.3.3. se puede observar un ejemplo de estructura de datos con sus respectivas propiedades ya clasificadas.

```
# Science
d_sciences = {
    'P183': 'endemic to',
    'P141': 'IUCN conservation status',
    'P2597': 'Gram staining',
    'P171': 'parent taxon',
    'P105': 'taxon rank',
    'P703': 'found in taxon',
    'P680': 'molecular function',
    'P682': 'biological process',
    'P681': 'cell component',
    'P2548': 'strand orientation',
    'P702': 'encoded by',
    'P688': 'encodes',
    'P128': 'regulates (molecular biology)',
    'P684': 'ortholog',
    'P1057': 'chromosome',
    'P196': 'minor planet group',
    'P566': 'basionym',
    'P59': 'constellation'
}
```

Fig. 6.2.3.3. Ejemplo de la estructura de datos creada para la propiedad Science. Fuente: Elaboración propia, 2018.

6.2.4. Cálculo de coeficientes

Para el cálculo de coeficientes, se sigue el proceso explicado en el octavo paso de la metodología. Este proceso consta de tres iteraciones para obtener un cálculo preciso de los coeficientes. Para cada una de las iteraciones es necesaria la implementación de un script.

En el primer script se calculan los coeficientes relacionados con las propiedades descriptivas, pero teniendo en cuenta para la ponderación las propiedades neutras. Para ello se recorre el volcado, y para cada ítem Q se cogen sus propiedades. Una vez se tienen las propiedades, se comprueba a que categoría pertenecen. Después se calculan los coeficientes en base a las ocurrencias de las propiedades en las estructuras de datos definidas, y por último se almacenan en base de datos. En la Fig. 6.2.4.1. se puede ver parte del código para el proceso del cálculo de coeficientes en la primera iteración.

```
        for snak in claimlist:
            # Mainsnak
            mainsnak = snak['mainsnak']
            # Property
            p_property = mainsnak['property']
            # Check if property references Q item
            if mainsnak['datavalue']['type'] == 'wikibase-entityid':
                # Insert into property list
                p_list.append(p_property)
    # Check if properties are into dictionaries
    for p_item in p_list:
        if p_item in pDictionary.d_agriculture:
            agriculture_counter = agriculture_counter + 1
            dictionary_occurrences = dictionary_occurrences + 1
        elif p_item in pDictionary.d_belief:
            belief_counter = belief_counter + 1
            dictionary_occurrences = dictionary_occurrences + 1
```

Fig. 6.2.4.1. Fragmento de código para cálculo de coeficientes en la primera iteración. Fuente: Elaboración propia, 2018.

En la segunda iteración, se calculan los coeficientes en base a las propiedades neutras, cogiendo para cada ítem referenciado por cada propiedad neutra los coeficientes calculados en la primera iteración. En las Fig. 6.2.4.2. y 6.2.4.3. se puede ver parte del código utilizado para esta segunda iteración.

```
if mainsnak['datavalue']['type'] == 'wikibase-entityid':
    if p_property in pDictionary.d_neutral:
        # Q2 custom_value
        q2item = 'Q{}'.format(mainsnak['datavalue']['value']['numeric-id'])
        # Select coefficients from step one
        sql_query = 'SELECT agriculture, belief, business, culture, education,
        # Get the tuple
        row = cursor.execute(sql_query).fetchone()
        # Check if exist
        if row:
            # Insert initial coefficients in initial dictionary
            initial_coef_dictionary[q2item] = row
```

Fig. 6.2.4.2. Fragmento de código para la obtención de coeficientes para ítems referenciados. Fuente: Elaboración propia, 2018.

```
# New dictionary to insert new coefficients
q2_coef_dictionary = defaultdict(list)
# Iterate the first dictionary
for key, value in initial_coef_dictionary.items():
    # Take the value into a list
    coef_list = list(value)
    # Iterate list
    for i in range(len(coef_list)):
        # Calculate new coefficient
        coef_list[i] = coef_list[i] / n_properties
    # Insert new key - value into new dictionary
    q2_coef_dictionary[key] = coef_list
```

Fig. 6.2.4.3. Fragmento de código para la ponderación de coeficientes para ítems referenciados. Fuente: Elaboración propia, 2018.

La tercera iteración del proceso es exactamente igual a la segunda, pero en esta ocasión se cogen los coeficientes calculados en la iteración número dos para cada uno de los ítems referenciados por propiedades clasificadas como neutras.

Después de esta tercera iteración, el valor de los coeficientes debe ser muy próximo a uno, lo cual ya sirve para pasar a la visualización de los resultados.

6.2.5. Visualización de la información clasificada

En este punto es necesaria la implementación de un conjunto de scripts para generar gráficos cogiendo los coeficientes calculados previamente. Para ello se usa una librería que ofrece Python para este cometido, Bokeh.

Todas las implementaciones relacionadas con este punto siguen un patrón común. Para todos los ítems, se coge la suma de cada uno de los coeficientes del paso anterior y se pondera por la suma total de todos para de esta manera obtener un valor normalizado a uno.

Una vez los valores a utilizar están normalizados, se ordenan de mayor a menos para determinar qué categorías tienen un porcentaje mayor de contenido. Con esta información se procede a la generación de los gráficos. En la Fig. 6.2.5.1. se puede ver un fragmento relevante de código asociado a uno de los scripts implementados, que puede ayudar a entender los resultados que se generan.

```
# Generated file
output_file("all_categories.html")
# Categories
categories = []
# Real coefficients
coefficient_list = []
# Iterate sorted dictionary
sorted_dictionary = list(sorted_dictionary)
for i in range(len(sorted_dictionary)):
    x = sorted_dictionary[i]
    key = x[0]
    value = float(x[1])
    categories.append(key + ' (' + str(value * 100) + ' %)')
    coefficient_list.append(value * 100)
# Create the figure
p = figure(x_range=categories, plot_height=500, title="Wikipedia All (%)",
           toolbar_location=None, tools="")
# Generate bars
p.vbar(x=categories, top=coefficient_list, width=0.6, color='blue')
```

Fig. 6.2.5.1. Fragmento de código para la generación de gráficos. Fuente: Elaboración propia, 2018.

7. Análisis de resultados

En base a los coeficientes obtenidos durante el proceso de cálculo de coeficientes y a la realización del proceso de visualización de la información clasificada, se obtienen distintos gráficos que proyectan los resultados del estudio realizado.

Destacar que el objetivo del estudio es analizar y visualizar el contenido de Wikipedia extraído de Wikidata clasificándolo en categorías definidas en este proyecto. Además, se generan resultados para tres versiones de Wikipedia: la versión en inglés, en español y en catalán.

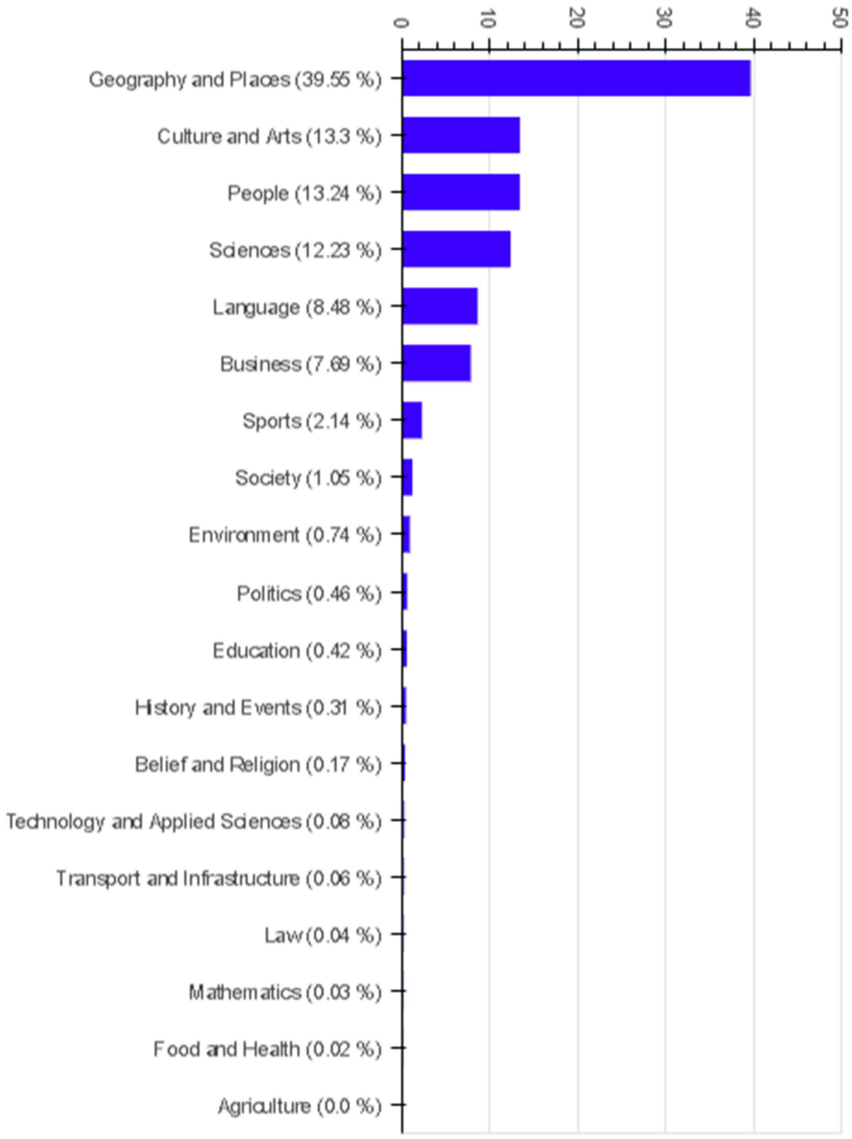

Fig. 7.1. Porcentajes de todo el contenido de Wikidata. Fuente: Elaboración propia, 2018.

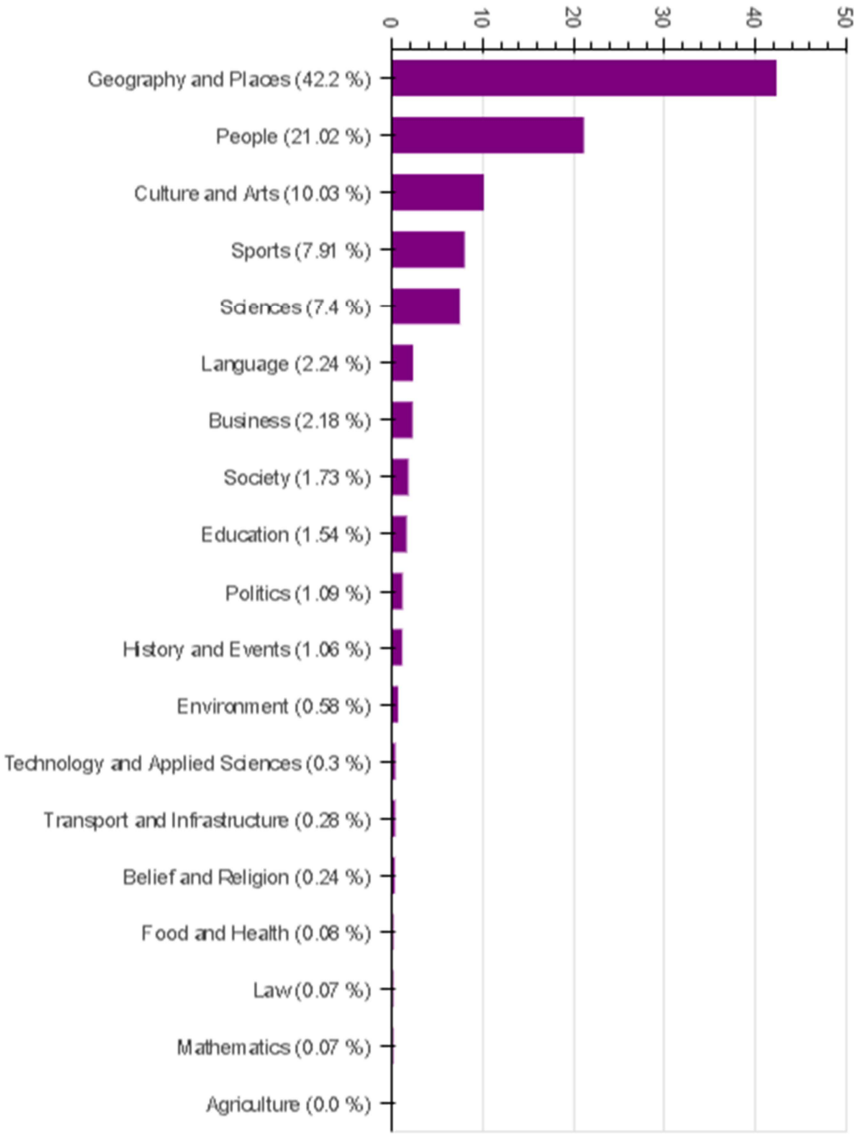

Fig. 7.2. Porcentajes de todo el contenido de la versión inglesa de Wikipedia. Fuente: Elaboración propia, 2018.

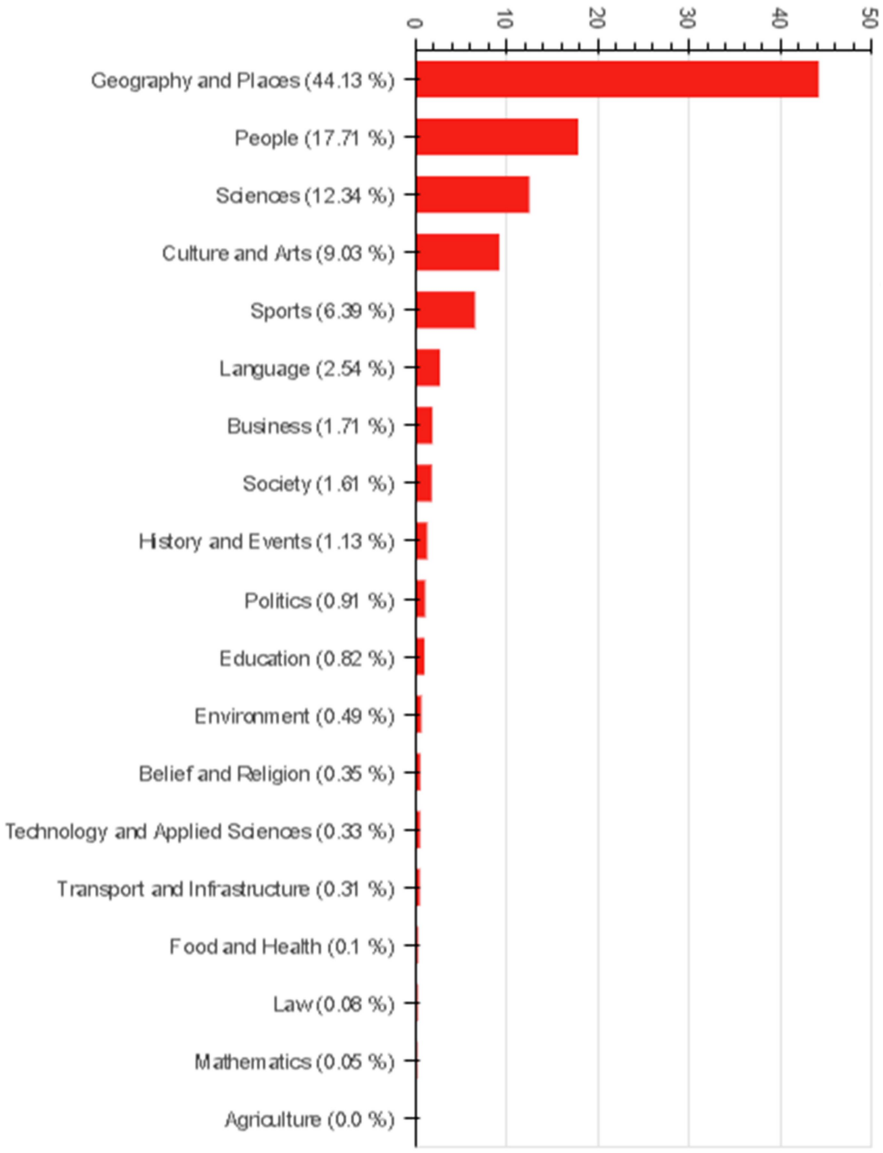

Fig. 7.3. Porcentajes de todo el contenido de la versión en español de Wikipedia. Fuente: Elaboración propia, 2018.

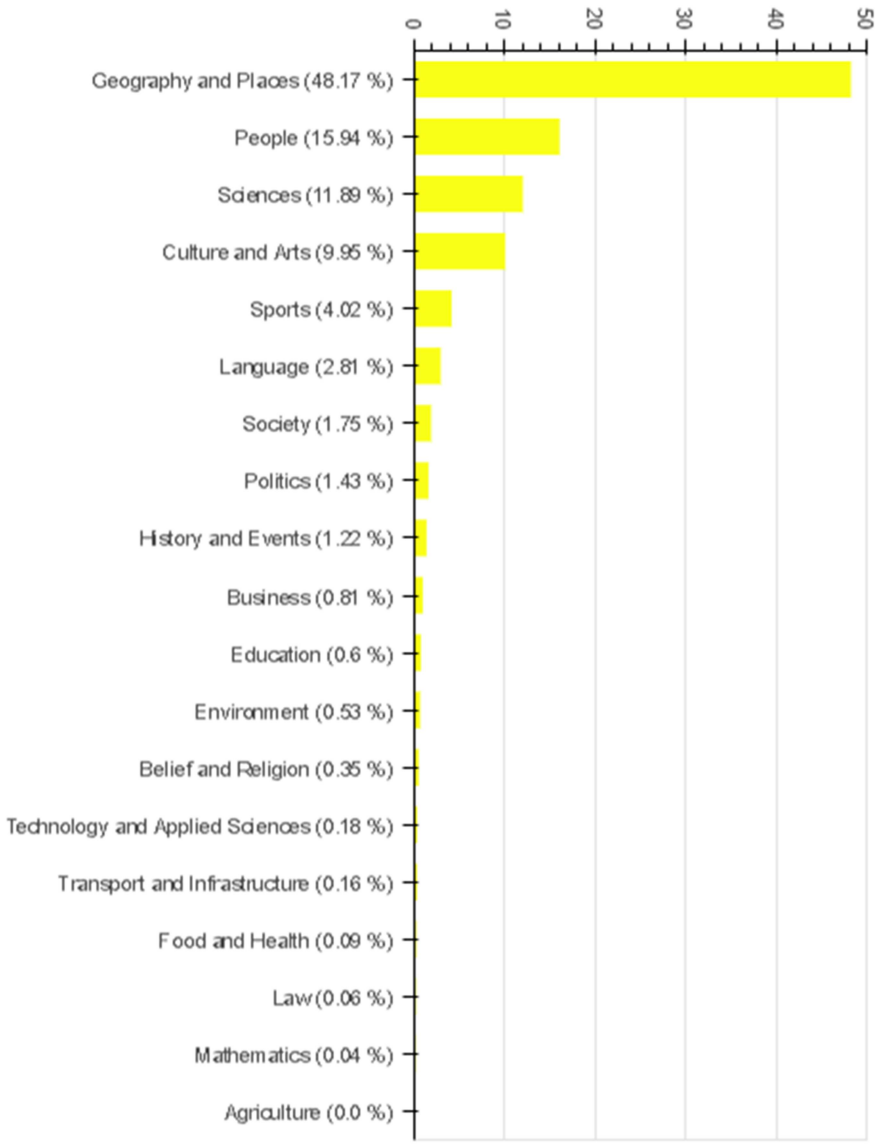

Fig. 7.4. Porcentajes de todo el contenido de la versión en catalán de Wikipedia. Fuente: Elaboración propia, 2018.

En la Fig. 7.1. se puede ver el gráfico que muestra los porcentajes de cada categoría para todo el conjunto de información existente en Wikidata. Se puede apreciar que la categoría con un porcentaje mayor de artículos en Wikipedia es la de *Geography and Places*, seguida por *Culture and Arts*, *People* y *Sciences*.

En la Fig. 7.2. se pueden ver los datos de la versión en inglés. Al igual que en la Fig. 7.1., *Geography and Places*, *Culture and Arts* y *People* siguen siendo las categorías predominantes.

En la Fig. 7.3. se puede ver el contenido clasificado de la versión de Wikipedia en español. Destacar que *Geography and Places* y *People* siguen siendo las categorías con más porcentaje de contenido, pero en este caso *Sciences* toma el relevo de *People*.

En la Fig. 7.4. se puede ver la información de las diferentes categorías en la versión en catalán. Destacar que las categorías con más información son las mismas que en la versión en español.

Haciendo una comparativa de los cuatro gráficos se aprecia que las categorías con un mayor porcentaje de información son similares en todos ellos. Las categorías *Geography and Places*, *Culture and Arts*, *People* y *Sciences* son las más destacadas. Siguen a estas categorías *Sports*, *Language* and *Business*, las cuales tienen un menos porcentaje de información pero también ocupan los primeros puestos en las versiones analizadas.

Se puede apreciar como las categorías que tienen menos información también son comunes en todos los gráficos analizados. *Food and Health*, *Law*, *Mathematics* y *Agriculture* ocupan los últimos lugares tanto en el gráfico que representa el contenido global como en las distintas versiones lingüísticas analizadas.

Cogiendo los resultados obtenidos y comparándolos con los estudios de Kittur y Farina se aprecia que las categorías con mayor porcentaje de información son comunes en los tres estudios realizados. Hay que tener en cuenta que este estudio se centra en el contenido a nivel global extrayendo los datos de Wikidata, mientras los dos estudios que se explican como referentes de este proyecto se centran en un análisis de la versión en inglés de Wikipedia.

Para finalizar el análisis, remarcar que tanto los resultados obtenidos durante el proyecto tanto los resultados de los estudios de Kittur y Farina tienen una fuerte dependencia al contenido clasificado en las categorías elegidas. En el caso de este proyecto, las categorías que poseen un mayor número de propiedades tienen un porcentaje mayor de contenido, ya que por ejemplo la categoría *Geography and Places*, entre las doscientas propiedades con más ocurrencias que se analizan, tiene treinta y nueve propiedades asociadas, mientras que la categoría *Agriculture* no tiene ninguna, razón por la cual se obtiene un cero por ciento de información en esta categoría.

8. Conclusiones

En la realización del proyecto se consiguen los dos objetivos principales establecidos.

Por un lado, se logra entender el funcionamiento de Wikipedia y Wikidata como plataforma de soporte al repositorio de información online más grande del mundo. Esto se consigue a través de la búsqueda de información realizada en el marco teórico y en el análisis de referentes de este proyecto. Destacar que en el caso de Wikidata, además se logra entender cómo está estructurada la información dentro de la plataforma y que elementos intervienen.

Por otro lado se consigue realizar un estudio de los contenidos a nivel global de Wikipedia usando datos e información procedentes de Wikidata. En este objetivo es muy importante el conjunto de categorías que se usan. A partir de los diferentes referentes analizados en este proyecto se extrae un conjunto de propiedades las cuales se usan para la clasificación del contenido de Wikidata asociado a Wikipedia.

En este punto es importante mencionar la importancia de la elección de las propiedades procedentes de Wikidata, las cuales se clasifican en las diferentes estructuras de datos utilizadas para su clasificación. Es determinante ya que su clasificación es muy importante, ya que a partir de ella salen los resultados del estudio.

También es importante que los resultados obtenidos se asemejen a los resultados de los dos estudios tomados como referentes, con sus respectivos matices.

Otro punto importante a destacar en este apartado es la utilización de tecnologías y herramientas *open-source*. Se demuestra que utilizando herramientas de libre acceso se pueden realizar estudios sobre Wikipedia y su ecosistema.

Como conclusión final, destacar la importancia de este estudio cuantitativo de información extraída de Wikidata, ya que cualquier estudio, proyecto o aporte que añada valor a Wikipedia y a su ecosistema ayuda a seguir mejorando esta gran plataforma, y la hace aún más grande.

9. Posibles ampliaciones

El estudio de los contenidos de Wikipedia extraídos de Wikidata aportado en este proyecto se puede ampliar en cinco puntos distintos a lo largo del proceso.

La primera ampliación viene por las categorías definidas para el estudio. Hay que tener en cuenta que para este proyecto se seleccionan diecinueve categorías generales para abarcar y cubrir todo el contenido, pero puede definirse un número más grande de categorías para obtener una mayor precisión en cuanto a temáticas de estudio.

La segunda de las ampliaciones se basa en el hecho de contemplar más categorías de Wikidata. Para este proyecto se seleccionan las doscientas primeras propiedades con más ocurrencias, que ya cubren más del noventa y nueve por ciento de las ocurrencias totales de propiedades que referencian a ítems. Pero si se quiere aumentar ese porcentaje se puede, seleccionando más categorías para clasificar en las distintas estructuras de datos definidas.

La tercera se centra en la fase de desarrollo. Puede haber un punto de mejora realizando más iteraciones para el cálculo de coeficientes, para así obtener unos resultados más precisos de lo que ya son.

La cuarta ampliación viene en la visualización de resultados. Al tener gran parte de la información que tiene Wikidata almacenada en base de datos propia, y teniendo la posibilidad de saber las preferencias temáticas de cada versión de Wikipedia, se pueden hacer más comparativas entre versiones de distintos países y ampliar el análisis del estudio.

La quinta ampliación es la transformación de este proyecto en un artículo científico. Esto no es una mejora en sí, pero la realización de un artículo de índole científico permite exponer este proyecto en las distintas conferencias relacionadas con el tema que se realizan durante el año en distintos países. Por ejemplo, se puede exponer en el Wikimedia Hackathon del año 2019, que se va a realizar en Praga, República Checa.

10. Bibliografía

[1] OpenMind [en línea] [consulta: 20 de Diciembre de 2017]. Disponible en https://www.bbvaopenmind.com/tim-berners-lee-y-el-origen-de-la-web/

[2] Evolución de la web [en línea] [consulta: 20 de Diciembre de 2017]. Disponible en http://profesores.elo.utfsm.cl/~tarredondo/info/networks/Evolucion_Web.pdf

[3] Wikipedia [en línea] [consulta: 10 de Enero de 2018]. Disponible en https://es.wikipedia.org/wiki/Web_2.0#Origen_del_t%C3%A9rmino

[4] Mehdi Khosrow-Pour, D.B.A: Encyclopedia of Information Science and Technology. Cuarta edición. ISBN: 9781522522553.

[5] Wikimedia [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://wikimediafoundation.org/wiki/Home

[6] Wiktionary [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wiktionary.org/wiki/Wikcionario:Portada

[7] Wikiquote [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wikiquote.org/wiki/Portada

[8] Wikibooks [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wikibooks.org/wiki/Portada

[9] Wikisource [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wikisource.org/wiki/Portada

[10] Wikispecies [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://species.wikimedia.org/wiki/Portada

[11] Wikinews [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wikinews.org/wiki/Portada

[12] Wikiversity [en línea] [consulta: 22 de Enero de 2018]. Disponible en https://es.wikiversity.org/wiki/Portada

[13] Wikivoyage [en línea] [consulta: 10 de Febrero de 2018]. Disponible en https://es.wikivoyage.org/wiki/P%C3%A1gina_principal

[14] Commons [en línea] [consulta: 11 de Febrero de 2018]. Disponible en https://commons.wikimedia.org/wiki/Main_Page

[15] MediaWiki [en línea] [consulta: 11 de Febrero de 2018]. Disponible en https://www.mediawiki.org/wiki/MediaWiki/es

[16] Meta-wiki [en línea] [consulta: 26 de Febrero de 2018]. Disponible en https://meta.wikimedia.org/wiki/Main_Page

[17] Incubator [en línea] [consulta: 26 de Febrero de 2018]. Disponible en https://incubator.wikimedia.org/wiki/Incubator:Main_Page/es

[18] Cloud Services [en línea] [consulta: 7 de Marzo de 2018]. Disponible en https://wikitech.wikimedia.org/wiki/Help:Cloud_Services_Introduction

[19] Creative Commons [en línea] [consulta: 7 de Marzo de 2018]. Disponible en https://creativecommons.org/licenses/by-sa/3.0/es/

[20] What is Wikipedia [en línea] [consulta: 17 de Marzo de 2018]. Disponible en https://upload.wikimedia.org/wikipedia/commons/e/e8/Wikipedia-leaflet-en.pdf

[21] Alexa [en línea] [consulta: 17 de Marzo de 2018]. Disponible en https://www.alexa.com/topsites

[22] Definicion.de [en línea] [consulta: 17 de Marzo de 2018]. Disponible en https://definicion.de/wikipedia

[23] Wikipedia [en línea] [consulta: 17 de Marzo de 2018]. Disponible en https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

[24] Wikidata [en línea] [consulta: 22 de Marzo de 2018]. Disponible en https://www.wikidata.org/wiki/Wikidata:Introduction

[25] CC Public Domain Dedication [en línea] [consulta: 24 de Marzo de 2018] Disponible en https://creativecommons.org/publicdomain/zero/1.0/

[26] What's in Wikipedia? [en línea] [consulta: 28 de Marzo de 2018]. Disponible en http://www.pensivepuffin.com/dwmcphd/syllabi/insc547_wi13/papers/wikipedia/kittur-socialcategory-CHI09.pdf

[27] Automatically assigning Wikipedia articles to macro-categories [en línea] [consulta: 28 de Marzo de 2018]. Disponible en http://airwiki.ws.dei.polimi.it/images/3/3e/Macro-categories.pdf

[28] SQLite [en línea] [consulta: 3 de Abril de 2018]. Disponible en https://www.sqlite.org/index.html

[29] Python [en línea] [consulta: 9 de Abril de 2018]. Disponible en https://www.python.org/

[30] Eclipse [en línea] [consulta: 15 de Abril de 2018] Disponible en https://www.eclipse.org/downloads/packages/eclipse-ide-java-developers/heliossr1

[31] PyDev [en línea] [consulta: 15 de Abril de 2018]. Disponible en http://www.pydev.org/

[32] Bokeh [en línea] [consulta: 15 de Abril de 2018]. Disponible en https://bokeh.pydata.org/en/latest/